# Getting in the Statistical Ballpark

## **Statistics is the study of how data is collected, organized, and interpreted**

Divisions of statistics

* Descriptive (🡪EDA)
* **Graphical**
* Numerical
* Inferential
* Sampling
* Hypothesis testing

Statistics tries to uncover **two key aspects** of all data:

the underlying **pattern** (**model**),

and **variability**

**Models**

A model is a simplified representation of physical reality based on a small number of suppositions or assumptions.

**Question:** What is the model of a coin flip that predicts ½ likelihood of either heads or tails?

**Models allow us to define relevant variables, anticipate relationships among them, and draw conclusions based on the observed relationships**

**Variability**

Any repeated measurement with sufficiently fine resolution will give varying values for the measured quantity. In science this variability is called *random error*, though the effect need not be random nor involve an error.

**Question:** What are some general causes of variability?

**Activity 1:** The “F” Test (Melton, *J. Stat. Ed.*, v. 12, n. 2, 2004)

You will be asked to count the number of letters “f” in a short paragraph. The distribution of answers will be collected as data. Prior to counting, create a model of what you expect to observe for the class. For the model, assume that the actual number of letters “f” is 10.

**My model** Min = Most likely = Max =

**My actual count** Num =

*Aggregated class model Aggregated class data*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of letters | Frequency |  |  | Number of letters | Frequency |
| 0 |  |  |  | 0 - 4 |  |
| 1 |  |  |  | 5 - 9 |  |
| 2 |  |  |  | 10 - 14 |  |
| 3 |  |  |  | 15 - 19 |  |
| 4 |  |  |  | 20 - 24 |  |
| 5 |  |  |  | 25 - 29 |  |
| 6 |  |  |  | 30 - 34 |  |
| 7 |  |  |  | 35 - 39 |  |
| 8 |  |  |  | 40 - 44 |  |
| 9 |  |  |  | 45 - 49 |  |
| 10 |  |  |  | 50 - 54 |  |
| 11 |  |  |  | 55 - 59 |  |
| 12 |  |  |  | 60 - 64 |  |
| 13 |  |  |  | 65 - 69 |  |
| 14 |  |  |  | 70 - 74 |  |
| >14 |  |  |  | 75 - |  |

**Activity 2:** The “Most” Threshold Test (Melton, ibid.)

When you think of the word “most,” what number (percentage) comes to mind? Write down on an index card a number between 0 and 100 that represents in your mind the smallest percentage that satisfies the description of “most.”

**Question:** Do you expect the class results to vary more or less than the “F” test results? Why?

**Types of Numerical Plots**

Plots are used to summarize data in a visual form. Just as “a picture is worth a thousand words,” plots can reduce complex data to a simpler form and expose the underlying patterns (model).

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Horizontal Axis** | **Vertical Axis** | **Comment** |
| Scatter (X-Y) plot | Real numbers (decimals, fractions, integers) | Real numbers | Plots one variable vs. another. Needs no special data preparation. |
| Line plot | Ordered real numbers, such as time | Smoothly changing real numbers | Works best with clearly related variables. X-axis data should be sorted. |
| Histogram | Ordered real numbers grouped into ranges or qualitative categorical data | Real numbers | X-axis data must be sorted and grouped. Vertical axis data is sum of all data in a given range or category. |
| Dot plot | Same as histogram | Integers | Same as histogram. Most useful for small amounts of data that need to be counted. |
| Box plot | None. Categorical data is used but not along axis. | Ordered real numbers | Data for each category must be sorted and grouped into quartiles (“five-number summary” technique) |

**Scatter (X-Y) Plot**

A scatter plot uses a marker (e.g. a dot) to show the value of the y-variable associated with each value of the x-variable. It works with all real-number data and requires no sorting or other data preparation. It is useful for comparing multiple variables with each other and confirming expected relationships (models) or finding new ones. Scatter plots are the primary graphical technique for the start of exploratory data analysis (EDA).

**Activity 3:** How does MLB winning % depend on offensive production?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Team** | **Winning %** |  |
| 1 | Atlanta Braves | 0.568 |  |
| 2 | Miami Marlins | 0.396 |  |
| 3 | New York Mets | 0.451 |  |
| 4 | Philadelphia Phillies | 0.500 |  |
| 5 | Washington Nationals | 0.505 |  |

<http://mlb.mlb.com>

(Plot winning % vs. team order) (Plot winning % vs. chosen variable)

Winning % seems to be most correlated with \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Histogram**

A histogram uses a column to represent the frequency or value of a quantity within a given category (quantitative or qualitative) or range of values (bin). (Technically, if categories are used it’s often called a **bar graph**.) The width of the column shows the bin size (in the case of real numbers) while the height indicates the value, which can be any real number. For real numbers along the x-axis, the numbers must be sorted (usually from smallest to largest) and grouped into appropriate bins. Bins defined by a given range always include the lower end of the range but exclude the upper end. For example, a bin of [10, 20] includes all values of x such that 10 ≤ x < 20. In the case of integers, this bin would include (10, ..., 19).

**Note:** A dot plot is a histogram with the columns replaced by discrete dots, making the value easier to count. This is only possible if the y-axis data is discrete (represented integers).

**A histogram is especially useful for gauging variability within a constant model.**

Example: What are the favorite colors of a class of 20 1st graders?

**Question:** What is the model assumed in this example?

**Histogram (Bar graph) Dot plot**

**Followup Question 1:** What color is *most likely* to be the favorite?

**The most likely quantity is referred to as the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.**

**Followup Question 2:** What colors are preferred by most children?

**Activity 4:** Analyzing the “F” test

Use one of the plots already discussed to compare the model of the “F” test to the actual result. Which plot type makes the most sense? Draw the plots below or do them on a computer. How do the modes and variability compare? Was the model accurate? Why or why not? Based on these results, what do you think is the correct number of f’s?

**Note:** To compare the model with the actual results, multiply the model’s x-values by 5. That is, replace each number with the appropriate 5-number bin (0 🡪 0 – 4, 1 🡪 5 – 9, etc.)

**Analyzing Data Distributions**

The goal of descriptive statistical analysis is to determine the underlying **pattern (model)** and the **variability**. The analysis is based on **4** characteristics of the data distribution. These characteristics only apply to *numerical* distributions, not to distributions over qualitative categories.

* Central tendency

**mode**, **median**, mean

* Spread

**range, interquartile range**, standard deviation

* Shape

**skewness**, kurtosis, **number of peaks** (modality)

* Unusual features

gaps, **outliers**

Determining the quantities that define these characteristics is a straightforward process.

**Five-Number Summary**

Any real-number distribution can be neatly summarized by **5** numbers: the minimum x-value, the maximum x-value, and the three quartile separators (Q1, Q2, and Q3). The procedure is

* Sort the distribution on the x-data from lowest to highest.
* Divide the data exactly in half. This means finding the x-value for which half the x-values are below and half are above. For an odd number of data points (n), the data is divided by the [(n + 1)/2]th data point. For an even number of data points (n), the data is divided by the x-value halfway between the [n/2]th and the [(n + 2)/2]th data points. This x-value is Q2.
* Divide each half in half again according to the preceding method. The lower divider is Q1 and the higher one is Q3.
* Summarize the distribution by (minimum, Q1, Q2, Q3, maximum)

**The five-number summary allows the 4 characteristics to be described quantitatively.**

**Median**—the x-value Q2

**Range**—equal to (maximum – minimum)

**Interquartile Range (IQR)**—equal to the central range (Q3 – Q1)

**Skewness**—skewed *right* means (Q3 – Q2) > (Q2 – Q1);

skewed *left* means (Q2 – Q1) > (Q3 – Q2)

**Outlier**—any data point xo more than 1.5 IQR’s from the two central quartiles

i.e., (Q1 – xo) > 1.5 IQR or (xo – Q3) > 1.5 IQR

**Example:** Data on MLB winning % before the 2013 All-Star break

30 teams 🡪 Q1, Q2, and Q3 are the 8th, halfway between 15th and 16th, and

23rd x-values.

Min = 0.351, Q1 = 0.453, Q2 = 0.500, Q3 = 0.558, Max = 0.613

Consequently, median = 0.500, IQR = 0.105, the distribution is nearly symmetric with a slight right skew, and there are no outliers.

**Box Plot**

A box plot is designed to provide a visual five-number summary of one or more data sets. The “box” displays the IQR with additional lines (“whiskers”) extending to the minimum and maximum. A box plot can be partially constructed on Excel using the second “stock” chart under the “Other Charts” tab. The data gets entered in a column as Q1, maximum, minimum, Q3. This is illustrated with the winning % data.

**Maximum**

**Q3**

**Median**

**Q1**

**Minimum**

**Activity 5:** Testing a Political Claim

Recently the Democratic Governors Association (DGA) sent out a press release on data from the Pew Charitable Trust, entitled “Republican-Led States Only States to Lose Jobs.” Use five-number summaries and box plots for Democratic and Republican data separately to test this claim.

(<http://democraticgovernors.org/republican-led-states-only-states-to-lose-jobs/>)

Is the claim correct?

What model is the DGA claiming?

Does the evidence support the model?

**References**

The “F” and “Most” tests are adapted from Kim I. Melton, *Journal of Statistics Education* Volume 12, Number 2 (2004).

<http://www.amstat.org/publications/jse/v12n2/melton.html>

A short but formal overview of descriptive statistics is available in S.S. Sonnad, “Describing Data: Statistical and Graphical Methods,” *Radiology* Volume 225, Number 3 (2002).

<http://radiology.rsna.org/content/225/3/622.full.pdf>

Tutorials on descriptive statistics are available at Stat Trek.

<http://www.stattrek.com/tutorials/statistics-tutorial.aspx>

Fully searchable baseball statistics are available at the Major League Baseball website.

<http://mlb.mlb.com/>

The press release ““Republican-Led States Only States to Lose Jobs” is from June 14, 2013 on the Democratic Governors Association (DGA) website. It includes a link to the original Pew Charitable Trust study.

<http://democraticgovernors.org/republican-led-states-only-states-to-lose-jobs/>